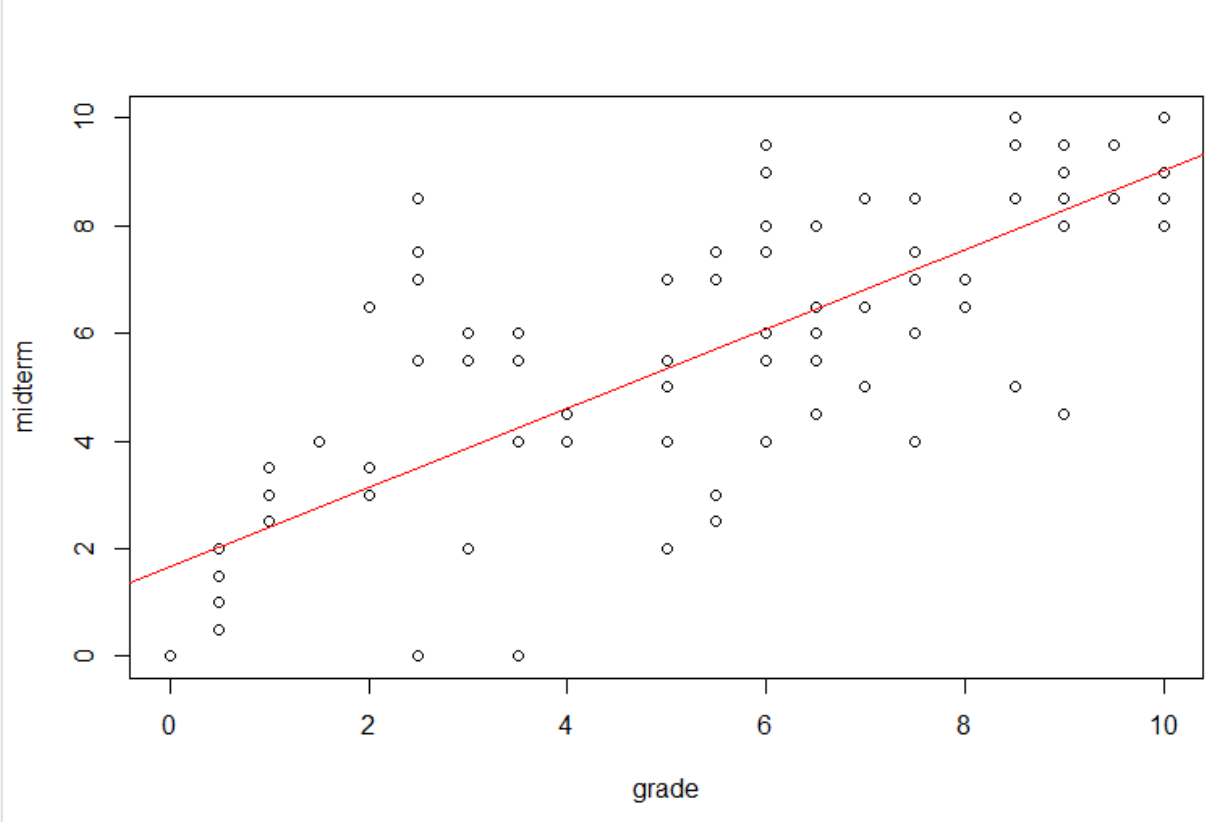
4η ΕΡΓΑΣΙΑ ΣΤΑΤΙΣΤΙΚΗ ΣΤΗΝ ΠΛΗΡΟΦΟΡΙΚΗ

ΠΛΑΤΩΝ ΚΑΡΑΓΕΩΡΓΗΣ 3180068

ΝΙΚΟΛΕΤΑ-ΚΛΕΙΩ ΠΑΤΑΤΣΗ 3180266

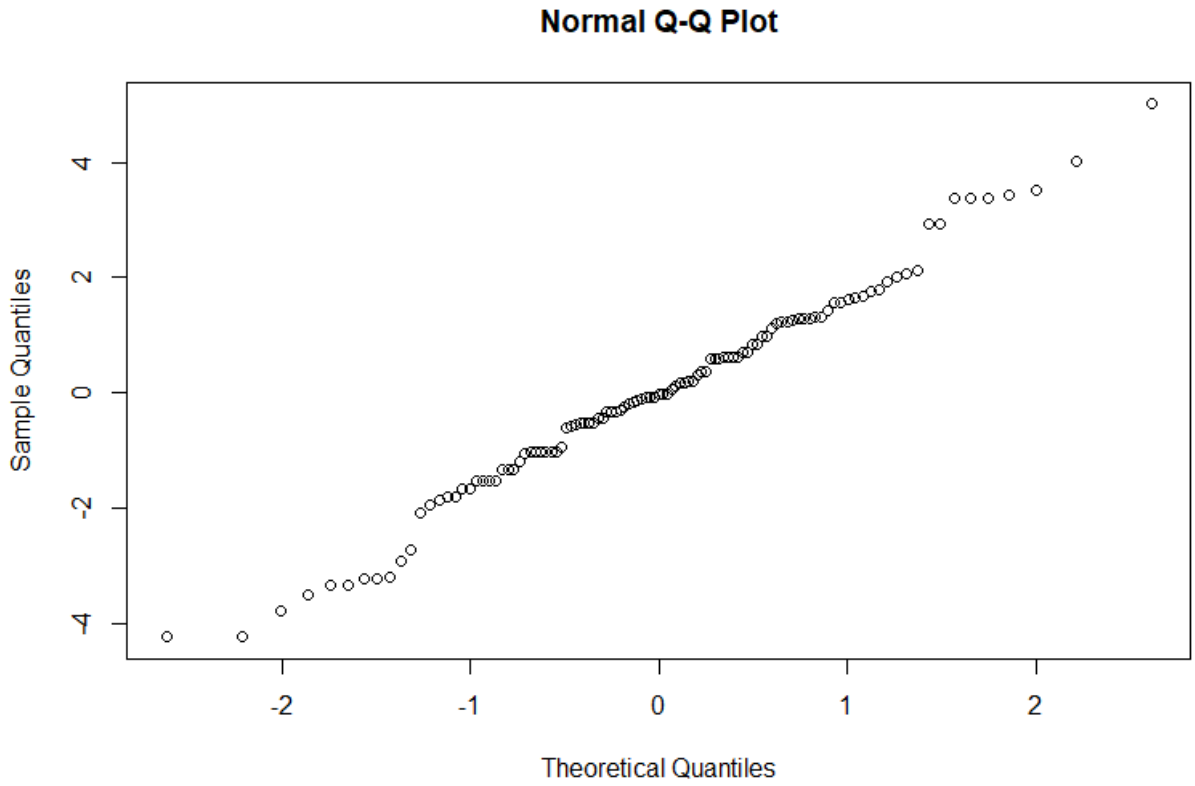
1)

α) Σε πρώτη φάση φορτώνουμε τα δεδομένα στην R. Έπειτα, τρέχουμε την εντολή plot(midterm~grade) και ακολούθως τις εντολές m<-lm(midterm~grade) και abline(m,col="red"). Το αποτέλεσμα φαίνεται παρακάτω.



Αρχικά, παρατηρούμε ότι η σχέση των δεδομένων είναι γραμμική καθώς δεν υπάρχουν σημαντικές αποκλίσεις. Επιπλέον, υπάρχει ομοσκεδαστικότητα καθώς οι τιμές αριστερά και δεξιά της γραμμής είναι ομοιόμορφα κατανεμημένες, χωρίς σημαντικές διαφορές. Τέλος, θα διερευνήσουμε για το αν τηρείται η κανονικότητα των τιμών midterm για

κάθε τιμή grade, χρησιμοποιώντας τις κατακόρυφες αποστάσεις, τα residuals. Εκτελούμε την εντολή qqplot(m$residuals)



Όπως βλέπουμε το αποτέλεσμα είναι πολύ κοντά στην κανονική κατανομή άρα μπορούμε να πούμε ότι τηρείται και η τελευταία απαίτηση. Επομένως τα δεδομένα είναι κατάλληλα και μπορούμε να κάνουμε εξαγωγή συμπερασμάτων από τα αποτελέσματα που θα προκύψουν.

β) Αρχικά με βάση την εκφώνηση πρέπει να αλλάξουμε τους άξονες x,y οπότε φτιάχνουμε νέο plot με την ίδια φιλοσοφία απλά με αντίστροφες τιμές. Τους συντελεστές β0, β1 τους υπολογίζουμε με την εντολή m<- lm(grade~midterm). Ο συντελεστής ευθείας β1 είναι ίσος με 0.8290192 ενώ η σταθερά β0 είναι ίση με 0.5756001. Τώρα θα υπολογίσουμε το 95% διάστημα εμπιστοσύνης.

```
> m$coefficients[2]->b1
> b1
  midterm
0.8290192
> SEb1<-summary(m)$coefficients[2,2]
> t<- -qt(0.025,df=109)
> b1 + c(-1,1)*t*SEb1
[1] 0.7035840 0.9544545
> 
```

Οι εντολές που εκτελούμε, υπολογίζουν τη μεταβλητή t γνωρίζοντας το βαθμό ελευθερίας από το summary(m) που εκτελέσαμε παρακάτω, το τυπικό λάθος SE και φυσικά την κλίση της ευθείας β1. Όπως βλέπουμε το 95% διάστημα εμπιστοσύνης που προκύπτει είναι 0.7035840 0.9544545.

γ) Η εντολή summary(m) για άλλη μία φορά μας δίνει την απάντηση.

```
> summary(m)

Call:
lm(formula = grade ~ midterm)

Residuals:
    Min      1Q  Median      3Q     Max 
-5.1223 -1.3191 -0.1223  1.1342  4.6938 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.57560    0.38438   1.497    0.137    
midterm      0.82902    0.06329  13.099   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.926 on 109 degrees of freedom
  (16 observations deleted due to missingness)
Multiple R-squared:  0.6115,    Adjusted R-squared:  0.608 
F-statistic: 171.6 on 1 and 109 DF,  p-value: < 2.2e-16
```

Έχουμε H0: β1=0 και HA: β1!=0. Ουσιαστικά η μηδενική υπόθεση μας δείχνει ότι εάν ισχύει δεν υπάρχει σχέση μεταξύ των δύο μεταβλητών. Όπως βλέπουμε το pValue που υπολογίζεται από το summary παραπάνω είναι 2e-16 δηλαδή πρακτικά μηδέν. Άρα απορρίπτουμε τη μηδενική υπόθεση και μπορούμε ουσιαστικά λόγω της πολύ μικρής τιμής του pValue, να αποδεχθούμε την HA και να πούμε με σχετική ασφάλεια ότι οι δύο μεταβλητές σχετίζονται.

δ) Θα χρησιμοποιήσουμε τη συνάρτηση predict.

```
> predict(m,newdata=data.frame(midterm=7),interval="confidence")
       fit      lwr      upr
1 6.378735 5.960928 6.796541
> |
```

Όπως παρατηρούμε, το 95 % διάστημα εμπιστοσύνης είναι 5.90928 6.796541 άρα περιμένουμε ο βαθμός του φοιτητή να κυμανθεί ανάμεσα σε αυτές τις τιμές. Το fit είναι η μέση εκτίμηση καθώς είναι το κέντρο του διαστήματος και αναμένουμε το αποτέλεσμα να πλησιάζει αυτή την τιμή.

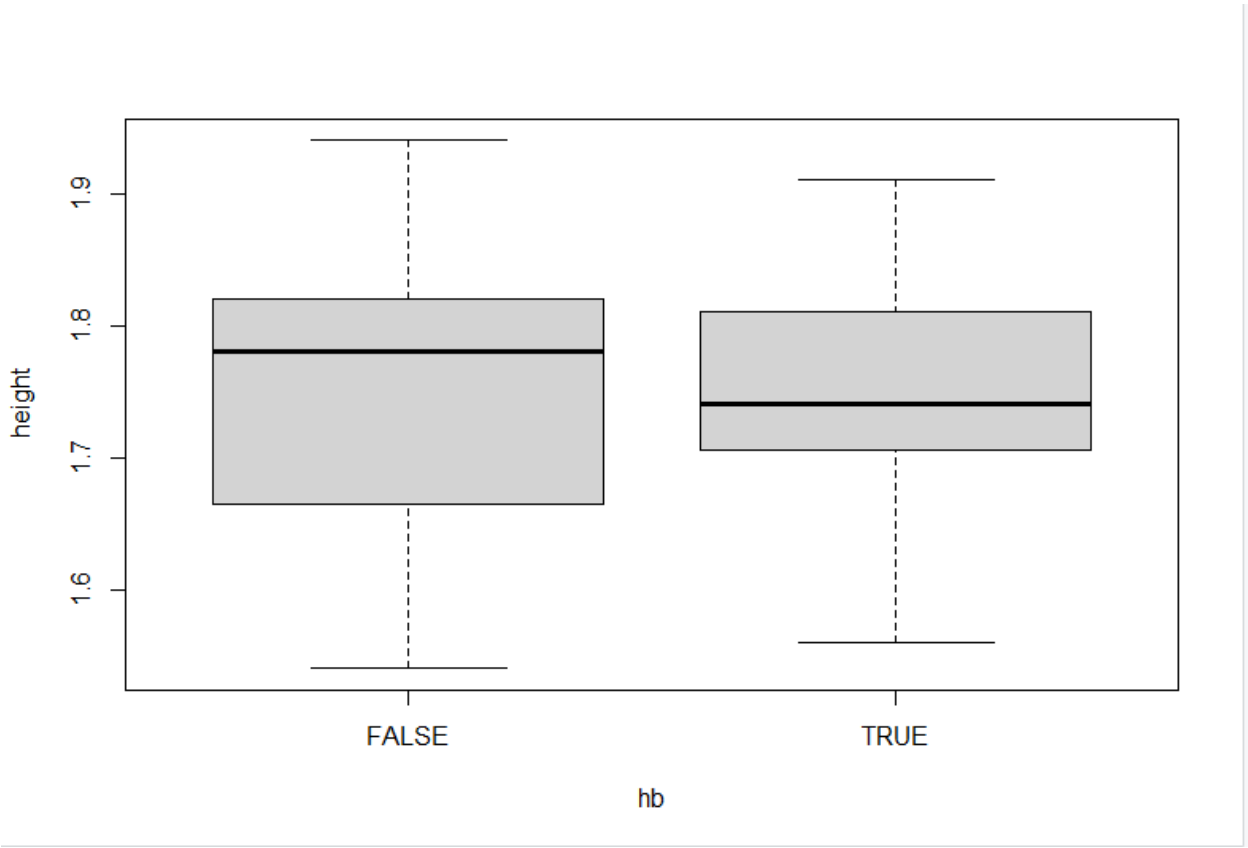
ε) Θα χρησιμοποιήσουμε και εδώ τη συνάρτηση predict.

```
> predict(m,newdata=data.frame(midterm=7),interval="predict")
       fit      lwr      upr
1 6.378735 2.537905 10.21956
> |
```

Βλέπουμε ότι το 95% διάστημα εμπιστοσύνης είναι 2.537805 10.21956 και είναι λογικό γιατί γίνεται μία πρόβλεψη επομένως αυξάνουμε το διάστημα πρόβλεψης για να μπορούμε να έχουμε μεγαλύτερη πιθανότητα να πέσουμε μέσα. Αντίστοιχα με πριν, το fit είναι η μέση εκτίμηση καθώς είναι το κέντρο του διαστήματος και αναμένουμε το αποτέλεσμα να πλησιάζει αυτή την τιμή.

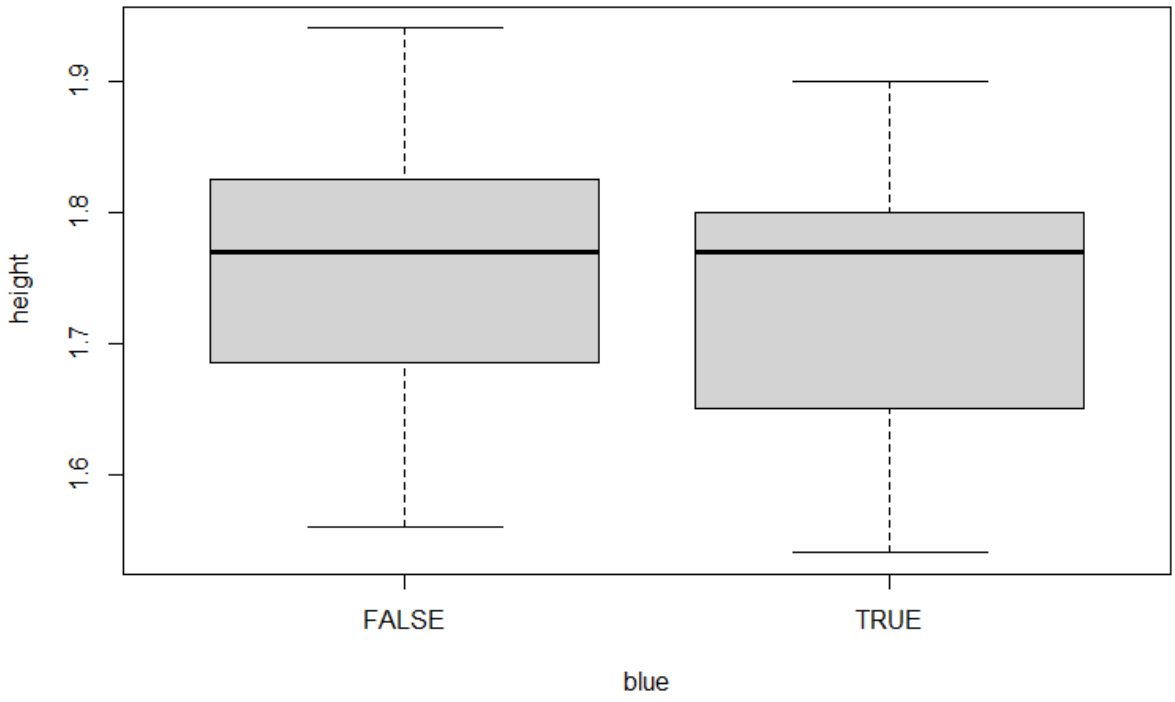
2)

α) Όπως έχουμε αναλύσει και σε προηγούμενες εργασίες, τα δεδομένα του ερωτηματολογίου προέρχονται από τυχαία δειγματοληψία και επομένως μπορούμε να χρησιμοποιήσουμε τύπους και να εξάγουμε συμπεράσματα. Αρχικά υπολογίζουμε ποιά είναι τα χρώματα που είναι πιο δημοφιλή. Τα πιο δημοφιλή χρώματα στο ερωτηματολόγιο είναι τα μαύρο, μπλε, μωβ. Το μωβ ήταν σε ισοβαθμία με το κόκκινο και το επιλέξαμε αυθαίρετα για την άσκηση. Θα ξεκινήσουμε τη σύγκριση για το μαύρο που είναι και το πιο δημοφιλές.



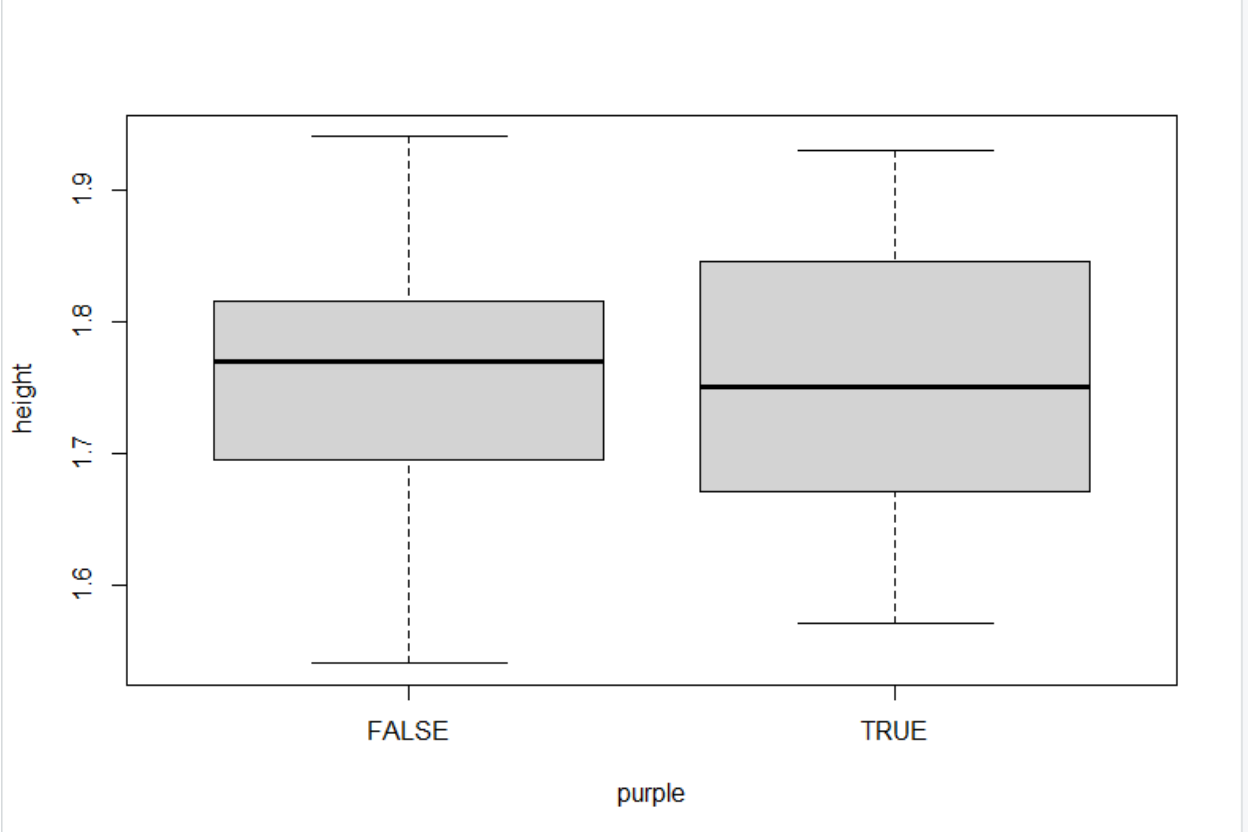
Όπως βλέπουμε υπάρχει μία απόκλιση στο median και στο range αλλά θεωρούμε ότι είναι σε φυσιολογικά επίπεδα και δεν θα λέγαμε πως το μαύρο χρώμα επηρεάζει το ύψος.

Στη συνέχεια εξετάζουμε το μπλε.



Εδώ υπάρχει μία πάρα πολύ μικρή απόκλιση στο median και το range αλλάζει ελάχιστα. Και εδώ θα λέγαμε πως δεν επηρεάζεται το ύψος ιδιαίτερα.

Τέλος, το μωβ.



Στο μωβ βλέπουμε ότι οι αποκλίσεις κυμαίνονται πάνω κάτω στο ίδιο επίπεδο με τις προηγούμενες περιπτώσεις, καταλήγουμε ότι ούτε εδώ επηρεάζεται το ύψος από το χρώμα.

β) Εδώ κάνουμε το εξής. Αρχικά, θα αναθέσουμε σε μία τοπική μεταβλητή το πλήθος των χρωμάτων που είναι μπλε, μαύρο ή μωβ. Έπειτα, θα αναθέσουμε σε μία άλλη τοπική μεταβλητή όλα τα ύψη που αντιστοιχούν σε ένα από αυτά τα τρία χρώματα. Μετά θα κάνουμε τον στατιστικό έλεγχο anova, για να δούμε κατά πόσο σχετίζονται τα χρώματα με το ύψος. Για να χρησιμοποιήσουμε anova, υποθέτουμε ότι τα δεδομένα μπορούν να χρησιμοποιηθούν και ότι είναι κατάλληλα, δηλαδή ότι οι τιμές του ύψους είναι κανονικά κατανεμημένες, σε σχέση με τις τιμές των χρωμάτων και ότι υπάρχει ομοσκεδαστικότητα. Εκτελούμε την εντολή aov(temp2~temp)->x όπου στα temp, temp2 έχουμε αυτά που αναφέραμε παραπάνω. Ακολούθως τρέχουμε την εντολή anova(x) και το αποτέλεσμα είναι το παρακάτω.

```
> aov(temp2~temp)-> x
> anova(x)
Analysis of Variance Table

Response: temp2
           Df  Sum Sq   Mean Sq F value Pr(>F)
temp        2 0.00571 0.0028536  0.3273  0.722
Residuals 66 0.57539 0.0087181
```

Όπως βλέπουμε, το pValue είναι 0.722 επομένως, δεν μπορούμε να απορρίψουμε την υπόθεση οι δύο αυτές μεταβλητές (ύψος-χρώμα) να σχετίζονται.

3)

α) Αρχικά για δική μας ευκολία δημιουργούμε μία μεταβλητή temp στην οποία θα βάζουμε τιμή TRUE=0 ή FALSE=1 ανάλογα με το αν ο βαθμός grade είναι μεγαλύτερος του 5. Το κάνουμε με την εντολή: temp<- ifelse(table$GRADE>5,0,1). Στη συνέχεια κάνουμε ένα scatterplot με την εντολή plot(midterm,temp). Η μεταβλητή απόκρισης μας θα είναι η μεταβλητή temp ενώ, η επεξηγηματική μας μεταβλητή θα είναι η midterm. Τώρα θα σχεδιάσουμε την καμπύλη της λογιστικής παλινδρόμησης.

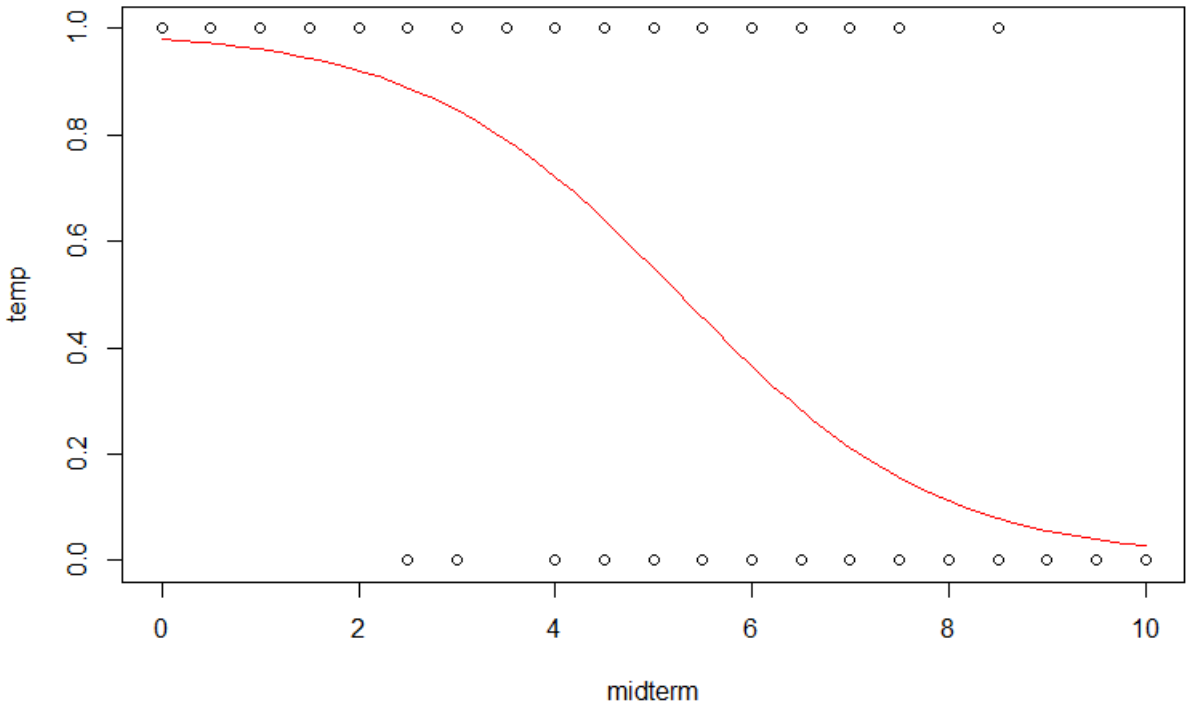
```
> glm(temp~midterm,family = binomial("logit"))

Call:  glm(formula = temp ~ midterm, family = binomial("logit"))

Coefficients:
(Intercept)      midterm
     3.9689      -0.7541

Degrees of Freedom: 110 Total (i.e. Null);  109 Residual
  (16 observations deleted due to missingness)
Null Deviance:      153.7
Residual Deviance: 90.57        AIC: 94.57
> glm(temp~midterm,family = binomial("logit"))-> m
> plot(midterm,temp)
> x = seq(from =0,to= 10,by=0.2)
> predict(m,newdata = data.frame(midterm=x),type="response")->y
> lines(x,y,col="red")
>
```

Με αυτές τις παραπάνω εντολές φτιάχνουμε το σχήμα που βλέπουμε παρακάτω.



Θεωρούμε ότι η προσέγγιση της λογιστικής παλινδρόμησης είναι πολύ καλή και επομένως πιστεύουμε ότι είναι κατάλληλη ως υπόδειγμα άρα αυτό μας δίνει το ελεύθερο να εφαρμόσουμε στην πορεία τους απαραίτητους στατιστικούς ελέγχους με ασφάλεια.

β) Πρώτα, τρέχουμε την εντολή, predict(m,newdata = data.frame(midterm=5),type="response") και το αποτέλεσμα της είναι 0.5494467. Αυτή η τιμή είναι σημείο στην καμπύλη η οποίο έχει την έννοια ποσοστού. Άρα, το αποτέλεσμα είναι αρκετά διχασμένο αφού είναι πολύ κοντά στο 50%. Ωστόσο έχει μία κλίση προς την αποτυχία στις εξετάσεις καθώς είναι πιο κοντά στις τιμές 1 όπου δείχνουν τους βαθμούς που είναι μικρότεροι του 5. Στο δια ταύτα το ποσοστό επιτυχίας είναι 46.1% όταν ένας φοιτητής γράφει 5 στην πρόοδο.

γ) Έστω H0: β1=0 και HA:β1!=0. Αν ισχύει η μηδενική υπόθεση τότε σημαίνει ότι οι δύο μεταβλητές που εξετάζουμε σχετίζονται. Έχοντας τρέξει την εντολή glm(temp~midterm,family = binomial("logit"))-> m από

πριν, τρέχουμε τώρα την summary(m) . Το αποτέλεσμα της το βλέπουμε παρακάτω:

```
> summary(m)

Call:
glm(formula = temp ~ midterm, family = binomial("logit"))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0981  -0.5828  -0.2354   0.5775   2.2470

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   3.9689     0.7856   5.052 4.37e-07 ***
midterm      -0.7541     0.1353  -5.575 2.47e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 153.65  on 110  degrees of freedom
Residual deviance:  90.57  on 109  degrees of freedom
  (16 observations deleted due to missingness)
AIC: 94.57

Number of Fisher Scoring iterations: 5
```

Παρατηρούμε ότι το pValue είναι 2.47e-08 επομένως μπορούμε με ασφάλεια να πούμε ότι απορρίπτεται η μηδενική υπόθεση και λόγω της πολύ μικρής τιμής του pValue μπορούμε να προσθέσουμε ότι θα αποδεχθούμε την εναλλακτική υπόθεση HA.

δ) Όπως είπαμε στο β ερώτημα που βασίζεται στο παραπάνω υπόδειγμα, το 46.1% των φοιτητών περνάει το μάθημα εάν γράψει 5 στην πρόοδο. Επομένως, εάν ένας φοιτητής γράψει 5 στην πρόοδο και είμαστε υποχρεωμένοι να βγάλουμε ένα συμπέρασμα για το αν πέρασε ή όχι τότε αυτό θα τείνει προς την αποτυχία του να περάσει το μάθημα.

4)

Ρίχνουμε το νόμισμα 100 φορές και έχουμε σαν αποτέλεσμα 44 κορώνες. Η πιθανότητα εμφάνισης κορώνας έχει να κάνει με το κατά πόσο το κέρμα είναι δίκαιο. Εάν το κέρμα είναι δίκαιο, τότε υπάρχει 0.5 πιθανότητα να βγει κορώνα ή γράμματα αντίστοιχα. Σε διαφορετική περίπτωση αν το κέρμα ευνοεί λίγο περισσότερο είτε την κορώνα είτε τα γράμματα τότε οι

πιθανότητες αλλάζουν. Φυσικά όμως εμείς δεν γνωρίζουμε αν είναι δίκαιο το κέρμα. Σύμφωνα με τη συνάρτηση της πιθανοφάνειας και με τη βοήθεια των διαφανειών, η μεγιστοποίηση της γίνεται όταν θ = 44/100 και άρα αυτή είναι η εκτίμηση της μέγιστης πιθανοφάνειας. Θέλουμε τώρα λοιπόν να υπολογίσουμε την πιθανότητα εμφάνισης κορώνας, βασισμένοι σε αυτή την συνάρτηση. Άρα, η πιθανότητα εμφάνισης δεν μπορεί να είναι άλλη από 44% καθώς η εκτίμηση της μέγιστης πιθανοφάνειας ισούται με το δειγματικό ποσοστό της τιμής που διερευνούμε, στην περίπτωση μας οι κορώνες.